Supplementary information

Supplementary Notes

Proteomics has many key advantages over genomics, especially in directly determining protein expression. Most quantitative experiments utilize a special tagging system, usually either chemically (ICAT, iTRAQ)[1] or biosynthetically (cells grown in media with one or more stable isotopes)[2]. Differences in protein expression are quantified by the relative intensity of conjoint spectra, each with a tag unit difference (for review[3]). Currently, retrospective comparisons between multiple datasets or with historical data are apparently impossible.

Label-free quantitative methods are just emerging[4-9] and growing in popularity, in part because they avoid the use of expensive labelling reagents, eliminate the extra analytical complexity from labelling, which requires additional ms/ms spectra interpretation, permit comparison of multiple datasets, and facilitate retrospective comparisons. These methods are based on using one mass spectrometry (MS) output feature of abundance, such as spectral or peptide counts[6, 7, 10, 11] as a means of determining relative protein abundance of the same protein in several samples. Spectral and peptide counts are easily extracted from MS data and show good correlation with protein abundance[10], but their "integer" or discrete nature may overestimate protein abundance, especially when low spectral counts are observed[4]. Chromatographic peak intensity and peak area have also been shown to correlate with protein abundance [12-15],[13, 14, 16], but require complex algorithms to integrate the area under the curve (AUC) or total elution curve for each isotope pattern. A number of tools can be used to extract peptide ion intensities following identification such as MSQuant (http://msquant.sourceforge.net)

(originally designed for quantifying stable isotope datasets) and Serac PeakExtractor[4]. Alternatively, publicly available tools such as ASAPratio[17], XPRESS[18] and RelEx[19] that have been designed specifically for the comparative quantification of stable-isotope labeled data using the extracted ion chromatogram method may also be modified and integrated into pipelines to compute intensities for specific peptide ions. The difficulty occurs in determining where one peak ends and a new peak begins, especially when signal and noise are not readily discriminated. Although this approach seems to work well, it requires high mass accuracy instrumentation and generally relies on "spiking" the sample with an exact amount of an internal standard[12, 20]. It also introduces materials into the sample before MS analysis, thereby creating a possible new bias from the standards as well as a new experimental variable external to the original dataset. Adding foreign materials to any sample is inherently worrisome and complicates universal application.

Several groups have added their own modifications to these approaches and these generally centre on the algorithm used in calculating the chromatographic intensity or a modification in how spectral counts are computed[21, 22]. Lu *et al* described a modified spectral counting technique, APEX, which improves on basic spectral counting methods by including a correction factor for each protein (called O*i* value) that accounts for variable peptide detection by MS techniques[23]. The technique is computationally challenging as it uses machine learning classification to derive peptide detection probabilities that are used to predict the number of tryptic peptides expected to be detected for one molecule of a particular protein. The predicted spectral count is compared to the proteins observed MS total spectral count during APEX computation of

protein abundances. The APEX Quantitative Proteomics Tool, later developed by Braisted *et al*[22] is an application that supports the technique.

Discovering differences in distinct biological samples based on a single shotgun measurement becomes futile simply because any second replicate "shotgun" MS measurement will identify 30-40% of proteins not found in another MS measurement of an identical sample[24]. This generation of partially overlapping datasets from identical samples suggests poor reproducibility of shotgun proteomic analysis which contributes to the impression of proteomics data "being soft" i.e., of dubious quality or lacking in stringency[25]. The statistical edict of "absence of evidence is not evidence of absence" holds true for shotgun proteomic data, as peptide detection and thus protein identification are dependent on a number of parameters intrinsic to the MS method itself, including peptide ionization efficiency. Ionization efficiency can be thought of as the tendency of the peptide to ionize and contribute to a mass spectrum. This is influenced mainly by the inherent structural properties of the peptide such as length, mass, amino acid composition and various biochemical properties such as hydrophobicity and pH under the experimental condition as well as the variable background present when the peptide elutes. Hence, higher stringencies including reaching 95% analytical completeness of a sample is required to identify meaningful protein differences between distinct biological samples; however, doing so requires 5-10 MS measurements of each distinct sample[24, 26].

Unfortunately, mass spectrometry measurements contain inherent biases and variations so that signals can frequently be corrupted by either systematic or even apparently random changes. Thus, replicate samples, regardless of the abundance feature used, will usually show variation in protein abundance which is likely not a reflection of

biological change. This highlights the need to normalize measurements in order to minimize inherent experimental bias and variability so that real changes in protein abundance between distinct samples can be reliably determined. Such statistical quantification becomes especially critical when multiple replicate samples are analyzed using single or multiple high-throughput, shotgun proteomics methods, which results in large volumes of data. In order for label-free quantitative proteomics to be more widely used, methods of normalizing and quantifying the data must be made available with full transparency and in a user friendly format.

Supplementary Data

Determining the magnitude of SI_N for any given change in protein abundance

The correlation between SI_N and the amount of protein loaded was R^2= 0.9239. The slope of the regression line is 1.223 with 95% CI of 1.101 to 1.345, meaning the magnitude of SI_N for any given change in protein abundance can be calculated. For example, a 2 fold change from 0.01 to 0.02 ng in reality produces a change of 0.01 but for SI_N this would be a change of 0.01223 (95% CI 0.01101 to 0.01345), whereas a 2 fold change from 100 to 200ng results in a SI_N increase of 122.3 (95% CI 110.1 to 134.5).

Analysis of the Standard protein mixture

The AUC, as calculated by each method outlined in the supplementary methods, was determined for each protein in the standard protein mixture and compared to the spectral count and SI_N values generated for the same proteins across the replicates. Only 16 of the 18 proteins were consistently detected in each of the 10 replicates, so these 16 common

proteins were compared across the replicates. As the actual amount of each protein in the standard mix was known, we chose to determine which method was best at accurately predicting the amount of the protein standard in the mix. Thus, the ug/ml predicated amount of each protein as determined by each quantitative method was compared to the actual loaded amount (ug/ml) originally added to the mix. The amount determined for each protein by each quantitative method was averaged across all replicates and compared to the actual loaded amount using ANOVA.

Most surprising was the Silva method which only managed to correctly determine the actual protein amount 18.75% of the time. This is very worrisome as this method (of normalizing the AUC of the proteins to the AUC of a spiked standard) is the most widely used AUC method in the literature. What was most interesting was that the spectral count method was better than the Silva method at accurately predicting the actual protein amount (37.5% vs 18.75%). In addition, this method consistently showed the largest variation across all of the proteins analyzed. Even the un-normalized AUC method showed less variation and accurately predicted the actual protein amount more often than this "normalization to internal standard method". Theoretically any of the 18 proteins in the mixture could be selected as the spiked standard as the exact amount of each protein is known. We arbitrarily chose ovalbumin, as the "standard" with which to normalize the data. To give AUC the best possible advantage and allow for all possible outcomes, we went back and checked the variation between the AUC (calculated the Silva way) calculated amount of all the individual standard proteins across the replicates and found that myoglobin had the smallest variation across replicates. Therefore, we re-analyzed the data using myoglobin as the "spiked standard" and thus all the proteins in the mix were

normalized by the AUC for myoglobin. Despite picking the best possible standard protein in the mix, this method only succeeded in correctly determining the correct amount of the protein in one additional case, meaning that it was only successful 4 times or 25% of the time (vs 3 times, 18.75%, using ovalbumin). These results clearly show that SI_N was substantially better than existing methods as a quantitative scoring procedure.

Testing distinct MS instruments

We clearly demonstrated the power of SI_N when comparing the replicate MS measurements of the same sample using the same MS method, regardless of initial sample load (Fig. 3). Therefore, we chose to determine if SI_N could be applied to MS datasets acquired by different MS instruments to facilitate their comparison (Supplementary Fig. 3). Proteins of endothelial cell plasma membranes isolated from lung were separated by SDS-PAGE followed by 1D-RP-LC-MS/MS analysis of all trypsin-digested gel slices using either an LCQ (1DLCQ) or an LTQ (1DLTQ) mass spectrometer, each with a distinct LC set up (see supplementary methods or Li *et al* [27] for details) The SI and SI_N values from the 769 proteins common between the two measurement types (3 replicates for each measurement) were averaged and plotted using the mean diamonds. As expected, the raw SI values show variation between the two MS measurement types (supplementary Fig. 3a) Application of SI_N normalizes the datasets so that no significant difference is detected between the different methodologies (supplementary Fig. 3b). A bivariate fit of the SI_N normalized 1DLCQ and 1DLTQ datasets indicates a strong positive linear correlation between the two datasets, which is confirmed by the Pearson's correlation of 0.796. The oval nature of the 95% C.I. density

ellipse also indicates the significant correlation between the datasets. Again, SI_N succeeded in controlling variation between MS measurements of the same sample acquired using different MS methodologies so that no significant difference can be detected between the datasets.

Comparison and quantification of proteins by different MS methodological analysis of the same sample

Using the lung P datasets, again generated from four different MS methodologies, we determined whether SI_N could be applied to different MS methods to facilitate comparison and quantification across all the datasets. We used all identified proteins, both common and distinct across all the datasets. Using GENESIS software, we first created a one-dimensional heat map (Supplementary. Fig. 7a) based on whether a protein was identified (black=No; Green=Yes), which clearly showed reproducibility within each method. After we applied the SI_N method to the same MS data to estimate protein levels, the 500 most abundant proteins from the datasets were presented in a heat map (Supplementary Fig. 7b). 2D LC-MS/MS (2DC) alone appeared the least sensitive method as it detected primarily highly abundant proteins, whereas first pre-fractionating with a gel before 2D-LC-MS/MS (G2DC) increased sensitivity and detected the widest range of proteins, even at very low levels. This clearly demonstrated the advantages to be gained by looking at SI_N values for each protein across each method, especially when compared to looking at a one dimensional, identified vs. non-identified approach (compare Supplementary Fig. 7a vs. 7b). Thus, the SI_N method also successfully revealed quantitative differences between methodologies.

Supplementary Discussion

Although various groups have looked at ms/ms ion intensities, to our knowledge, there are no reports linking the intensity of the ms/ms fragments to the abundance of the precursor ion and thus the abundance of the identifying peptides and protein. Tabb *et al* [28, 29] carried out extensive analysis on ms/ms fragment ions in terms of identifying trends in fragment ion peak intensity in the context of chemical composition, ion series and fragment mass. Analysis of multiple ms/ms spectra revealed that a significant number of identified ions are due to noise peaks. It is for this reason that we do not use the total ion chromatogram (TIC) for the spectrum when calculating our SI. Instead we only used the intensity of identifying peaks that match the precursor. Tabb *et al* demonstrated that fragment ions containing basic residues produce more intense peaks that those without basic residues (showed via analysis using Proteinase K, which produces a greater diversity of basic residue content in peptides). Tryptic peptides fragment in ion trap tandem MS producing prominent C-terminal y series ions and N-terminal b series ions. When basic residues (Lysine and Arginine) are at the N-terminus the b series is most intense, when basic residues are at the C-terminus, the y series is most intense. Therefore these basic tryptic peptides are most likely the more intense peaks in the spectra and this reduces the risk of including noise in the intensity calculation [28, 29]. In addition, Tabb *et al* has reported the incorporation of intensity values from ms/ms spectra to enhance peptide identification scores, because these intensity values reflect the basic residue content of the fragment ion, thus facilitating the generation of more accurate theoretical spectra [30, 31]. Venable *et al* [32] described an automated approach for the analysis of complex mixtures from tandem mass spectra. They adapted their RelEx program to extract and integrate ion

chromatograms from ms/ms scans for isobaric labeling strategies, such as ^{15}N labeling used in their paper. Therefore, in retrospect, it appeared logical, and subsequently rewarding, to investigate the use of ms/ms ions as an abundance feature.

Supplementary Methods

Sample preparation:

Sprague-Dawley female rats (150−250g; Charles River Laboratories) were used unless otherwise indicated, and all animal procedures were carried out in accordance with the Sidney Kimmel Cancer Center committee on Animal usage and Care (IACUC) standards. As described previously[33, 34], luminal vascular endothelial cell plasma membranes were directly isolated from rat lung and liver tissues with quality control showing ≥ 20-fold enrichment for known endothelial makers and ≥ 20-fold depletion of markers of other cell types and subcellular organelles. Sample purity was assessed with multiple antibodies against protein markers for endothelial membrane and other cellular compartments.

Western blot analysis

All antibodies were purchased commercially or obtained as gifts from other researchers. Custom polyclonal antibodies were provided by BioSource (Hopkinton, MA) and 21st Century Biochemicals (Marlboro, MA). Western blotting was carried out as described previously[33, 34]. Densitometry analysis was carried out using Scion Image software for PCs.

Mass spectrometry analysis

Gel Pre-fractionation: Proteins were pre-fractionated on SDS-PAGE gels prior to 2D-LC-MS/MS and Reverse Phase-MS/MS. Briefly, proteins in the samples were separated by SDS-PAGE (PAGEr gel, 8 - 16% T, 10 x 10 cm Cambrex Bio Science, Rockland, Inc. ME, USA) and visualized with colloidal Coomassie Blue staining (Invitrogen, Carlsbad, CA, USA.). Gel lanes were cut into slices (usually 70 but depending on sample and experiment, always ≥ 50) for in-gel proteolytic digestions. For RP-MS/MS, digested peptides were extracted from each gel slice three times with 20% ACN and 10% formic acid solution. The extracted peptide fractions were lyophilized. For 2D-LC-MS/MS, peptides extracted from each gel slice were first pooled into 7 groups then lyophilized.

Reverse-phase LC-MS/MS: For analysis by LCQ, lyophilized peptides were resuspended in 10 µl of buffer A (0.1% formic acid, 5% Acetonitrile (ACN)), and loaded onto a manually packed C18 microcapillary column under a Helium Pressure Cell, with approx. 600 psi. The bound peptides were eluted with 5 to 80% ACN gradients containing 0.1% formic acid over a 60- minute period. The eluted peptides were directly introduced into LCQ DecaXP (Thermo Fisher Scientific, Inc., Waltham, MA, USA) equipped with ESI nanospray ion source (Micro Sprayer, Mass Evolution, TX, USA). The flow rate was maintained at 200 to 250nl/min.

For analysis by LTQ, the lyophilized peptides were resuspended in 5 µl of buffer A and injected into a 5 mm trap cartridge (Dionex Corporation, Sunnyvale, CA, USA) for desalting using a FAMOS autosampler and a Switchos II system (Dionex Corporation, Sunnyvale, CA, USA). The desalted peptides were then back-eluted onto the analytical column, PepMap 100, C18, for the separation steps. The bound peptides were separated by a 110 minute ACN gradient (5% to 80% containing 0.1% formic acid) and directly

introduced into LTQ equipped with Nanospray I ion source (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The flow rate was maintained at 200 to 250nl/min.

Data acquisition from both the LCQ and LTQ was carried out in data-dependent mode. Full MS scan were recorded on the eluting peptides over the 400-1400 m/z range with one MS scan followed by three MS/MS scans of the most abundant ions. The temperature of the ion transfer tube of both mass spectrometers was set at 180ºC and the spray voltage was 2.0 kv. The normalized collision energy was set at 35% for both LCQ and LTQ. A dynamic exclusion was applied for Repeat Count of 2, a Repeat Duration of 0.5 minute, and an Exclusion Duration of 3 min.

2D- LC-MS/MS: Lyophilized peptides were resuspended with 30 µl of buffer A, then loaded onto a two-dimensional microcapillary column (manually packed C_{18} reversed phase and strong cation exchange column). The loaded samples were directly introduced into the LCQ mass spectrometer equipped with ESI nanospray ion source by eluting the bound peptides with a 2D-LC/MS/MS scheme controlled by Agilent 1100 HPLC quaternary pump[27]s. Briefly, 17 salt steps (ammonium acetate) were applied. Each salt step was followed by a 5 to 80% ACN gradient containing 0.1% formic acid to elute the peptides on the C_{18} column. The flow rate was maintained at 200 to 250nl/min. Data acquisition and the instrument setup were the same as the reversed-phase analysis, except the dynamic exclusion window was applied for duration of 10 minutes.

Gel-free 2D- LC-MS/MS: (2DC) analysis was performed as previously described [24].

Spiking experiments:

A protein standard mixture was prepared using protein standards (obtained from Sigma) added at various concentrations spanning a wide dynamic range.

Protein name	Uniprot accession number	Amount (fmol)	Amount loaded (ng)
Interleukin	P10145	0.5	0.0010
Fatty acid-binding protein	P05413	0.5	0.0018
C-reactive protein	P02741	0.5	0.0028
Annexin A5	P08758	0.5	0.00447
Glutathione S-transferase P	P09211	5	0.0290
Cathepsin G	P08311	5	0.0334
Insulin-like growth factor II	P01344	50	0.0933
Gelsolin	P06396	5	0.1037
Glutathione S-transferase A1	P08263	50	0.3185
Cathepsin D	P07339	50	0.3338
Antithrombin-III	P01008	50	0.6129
Alpha-lactalbumin	P00709	500	1.7587
Creatine kinase M-type	P06732	500	5.3837
Epidermal growth factor	P01133	5000	7.763
NAD(P)H dehydrogenase [quinone] 1	P15559	5000	38.73
Catalase	P04040	5000	74.478
Carbonic anhydrase 1	P00915	50000	359.335
Carbonic anhydrase 2	P00918	50000	363.687
Serum albumin	P02768	50000	829.91

The protein mixture was run on 1D-PAGE and stopped before they were separated. The bands containing the 19 proteins were digested with trypsin, and analyzed by Reverse-phase LC-MS/MS and 2D- LC-MS/MS as described above (n=4). Raw files were searched against a protein database containing sequences for the 19 spiked proteins as well as an additional 20 "decoy" proteins, which were not added to the mixture. Database search and identification parameters were as described below except 1 protein hits were also considered.

Database search:

The acquired MS/MS spectra were converted into mass lists using the Extract_msn program from Xcalibur and searched against a protein database containing human (97,361 entries), rat (52,881 entries) and mouse (112,998 entries) sequences using the Sequest program in the Bioworks^{TM} 3.1 for Linux (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The database includes protein entries from NCBI RefSeq and

SwissProt databases and was downloaded in April 2006 (Total entries, 262,200). The searches were performed allowing for tryptic peptides only with peptide mass tolerance of 1.5 Da for LTQ data, 2.0 Da for LCQ data, and a minimum of 21 fragmented ions in one MS/MS scan. Accepted peptide identification was based on a minimum ΔCn score of 0.1; minimum cross correlation score of 1.8(z=1), 2.5(z=2), 3.5(z=3). The peptides identified using these criteria showed much lower mass errors compared with other Sequest scores ([27] supplementary info). False positive identification rate was determined by the ratio of number of peptides found only in the reversed database to the total number of peptides found in both forward and reverse databases. The false positive identification rates were ≤ 1%. The positive protein identification results were extracted from Sequest.out files, filtered and grouped with DTASelect software using above criteria. Proteins were identified based on 2 unique significantly identified peptides.

In general, fragment ion intensity, peptide number and spectral counts were extracted from the DTAselect output files using a script written in-house (supplementary data). For the purpose of this manuscript, fragment ion intensity is defined as the total intensity of all detected fragment ions (ms/ms spectra) aligned with a specific peptide. The fragment ion intensity of each peptide that passes the threshold for identification that gives rise to a significantly identified protein (see above) is summed. The combination of these summed fragment ion intensities from all ms/ms spectra and peptides relating to this protein is combined and is referred to as the spectral index (SI) for that protein. For faster data acquisition, we used centroid algorithms for all of the MS analysis. In general, the centroid algorithms will sum the intensities if the ions have very close values, i.e.

isotope clusters. Therefore, the fragment ion intensities obtained are those that are recorded in Bioworks at the time of data acquisition.

Because DTAselect does not extract area under the curve (AUC) measurements (Bioworks Version 3.1 does not have the features for AUC calculation), we had to re-search our MS data to permit comparison of AUC and SI quantification. We searched one 2D-LC-MS/MS run from 4 replicate measurements of a gel section of our liver endothelial cell plasma membrane samples using Sequest on the cluster version of Bioworks 3.2. AUC values for each peptide were manually extracted using the AUC feature of Bioworks 3.2, and compared to SI values from the same gel section. The default parameters for an LTQ MS were used to calculate the AUC (see below).

The "standard protein mixture" experiments

Raw data files from 10 replicate analysis of the standard protein mixture [35] carried out by an LTQ mass spectrometer, were downloaded from the ISB website. We chose these datasets because it is the same mass spectrometer as we use in our own lab, and thus have all the software necessary for searching the data and extracting the required information. We searched the data against the same databases highlighted in the original paper using Sequest with Bioworks 3.2. The resulting data was sorted and group as described above using DTAselect for the calculation of spectral count and SI_N values. We used the peak area calculation function in Bioworks 3.2 (incorporates the ICIS algorithm) to calculate the AUC for each significantly identified peptide that was matched to a standard protein in the sample. We used the default parameters for the AUC as follows: mass tolerance 1.5amu, 5 point smoothing, minimum threshold for peak integration is 50,000. The AUC

for each protein was presented multiple ways, including methods corresponding to normalization approaches for AUC published in the literature. As SI_N is a normalized index, we thought it only fair to present the AUC data before and after it has been normalized by various published methods. These include:

1) total AUC: the AUC for each protein is presented as the sum of AUCs for all significantly identified peptides identifying each protein in the run (un-normalized data).
2) PA: this corresponds to the percentage peak area (PA), which is the default "normalization" in the Bioworks program, where the total AUC for a protein is expressed as a percentage of the total AUC for all identified proteins.
3) *Silva et al*: the average AUC for the 3 most intense peptides per protein is calculated and then normalized by the AUC of a protein standard.[20]. This approach is very similar to Mann *et al* [36], also very similar to other popular AUC methods that normalize to the AUC of a spiked internal standard.
4) *Old et al*: sum the AUC for each peptide, then each peptide is corrected by dividing the peptide by the sum of all peptide intensities. Similar peptides are compared across replicates and average peptide ratios are generated to reflect protein abundance[4].

The AUC, as calculated by each method outlined above, was determined for each protein in the standard protein mixture and compared to the spectral count and SI_N values generated for the same proteins across the replicates. Only 16 of the 18 proteins were consistently detected in each of the 10 replicates, so these 16 common proteins were compared across the replicates. The amount determined for each protein by each

quantitative method was averaged across all replicates and compared to the actual loaded amount using ANOVA. The mean value of each protein was compared to the actual loaded amount using the Tukey-Kramer HSD method[37, 38] (same principle as the t-test, but corrects for multiple testing).

Statistics

Datasets distribution: skewness and kurtosis: The skewness is a measure of distributions symmetry. For symmetrical distributions the skewness = 0, for right and left tailed distributions the skewness is >0 and <0, respectively. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Datasets with high kurtosis (>0) tend to have a distinct peak near the mean, decline rather rapidly, and have "heavy" tails. Data sets with low kurtosis (<0) tend to have a flat top near the mean rather than a sharp peak. Kurtosis = 0 for a normal distribution.

Unsupervised hierarchical clustering: Cluster analysis was performed on a dataset from 5 replicate MS measurements of endothelial cell plasma membranes isolated from kidney and heart samples using JMP 5.1, and using Wards hierarchical method [39].Ward's method is a hierarchical method designed to optimize the minimum variance within clusters (minimizes within-group dispersions). The clustering was unsupervised meaning without labeled classes, optimization criterion, feedback signal, or any other information beyond the raw data. Simply, we did not differentiate in any way the heart samples from the kidney samples. A Two-way clustering was performed, which is a data mining technique that allows simultaneous clustering of the rows and columns of a matrix [40-42].

Perl script to extract intensity values from DTAselect results file for SI_N calculation

```
#!/usr/bin/perl

$debug = 0;

my $infile = &get("Full path for DTASelect-filter.txt");
open(INFILE, $infile) || die "can't open file '$infile'";

my $outfile = &get("Output file");
open(OUTFILE, ">$outfile") || die "can't open file '$outfile'";

my $intensity_col;
print OUTFILE "id\tsum\n";
while (<INFILE>) {
  s/\s+$//;   # remove trailing newline
  my @line = split /\t/;
  if ($line[0] eq "Unique" and $line[1] eq "FileName") {
     for my $i (0 .. $#line) {
        $intensity_col = $i if $line[$i] eq 'TotalIntensity';
     }
     $mode = 1;
     next;
  }
  last if $line[0] eq "Unfiltered";
  last if $line[1] eq "Proteins";

  next if $mode != 1;
  if ($line[1] > 0) {
     &print;
warn "$line[1] is > 0 : $line[1]\n" if $debug;
     push(@names, $line[0]);
warn "adding $line[0] to \@names\n" if $debug;
  }
  else {
     die "count not find TotalIntensity column" if $intensity_col eq '';
warn "$line[1] is <= 0, adding $line[$intensity_col] for @names\n" if $debug;
     push(@values, $line[$intensity_col]);
  }
}
&print;

sub print {
  return if @names == 0;
  return if @values == 0;
  my $sum = 0;
```

```
  grep($sum += $_, @values);
  for my $name (@names) {
        print OUTFILE "$name\t$sum\n";
  }
  @names = ();
  @values = ();
}

sub get {
     local($prompt, $default) = @_;
     if ($default ne "") {
          &get2("$prompt [$default]: ", $default);
     }
     else {
          &get2("$prompt: ", $default);
     }
}

sub get2 {
     local($prompt, $default) = @_;
     local($mode) = $|;
     $| = 1;
     print "$prompt";
     $| = $mode;
     local($tmp);
     $tmp = <STDIN>;
     $tmp =~ s/\n$//;              # remove trailing \n
     $tmp = $default if $tmp eq '';
     return $tmp;
}
```

Supplementary References

1. Shiio, Y. *et al.* Quantitative proteomic analysis of Myc oncoprotein function. *Embo J* 21, 5088-5096 (2002).
2. Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1, 376-386 (2002).
3. Chen, X. *et al.* Amino acid-coded tagging approaches in quantitative proteomics. *Expert Rev Proteomics* 4, 25-37 (2007).
4. Old, W.M. *et al.* Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* 4, 1487-1502 (2005).

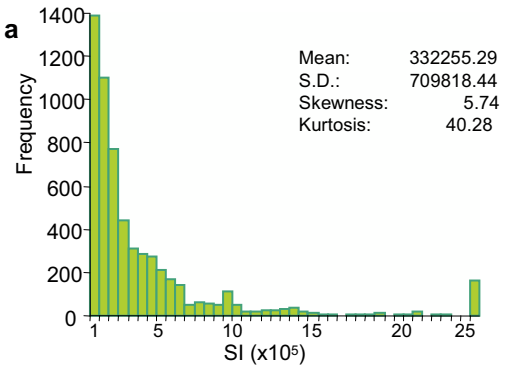
5. Zhang, B. *et al.* Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 5, 2909-2918 (2006).
6. Andersen, J.S. *et al.* Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 426, 570-574 (2003).
7. Gilchrist, A. *et al.* Quantitative proteomics analysis of the secretory pathway. *Cell* 127, 1265-1281 (2006).
8. Takamori, S. *et al.* Molecular anatomy of a trafficking organelle. *Cell* 127, 831-846 (2006).
9. Kislinger, T. *et al.* Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* 125, 173-186 (2006).
10. Liu, H., Sadygov, R.G. & Yates, J.R., 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 76, 4193-4201 (2004).
11. Ishihama, Y. *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 4, 1265-1272 (2005).
12. Cutillas, P.R. & Vanhaesebroeck, B. Quantitative profile of five murine core proteomes using label-free functional proteomics. *Mol Cell Proteomics* (2007).
13. Bondarenko, P.V., Chelius, D. & Shaler, T.A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* 74, 4741-4749 (2002).
14. Chelius, D. & Bondarenko, P.V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* 1, 317-323 (2002).
15. Silva, J.C. *et al.* Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. *Mol Cell Proteomics* 5, 589-607 (2006).
16. Gao, B.B., Stuart, L. & Feener, E.P. Label-free quantitative analysis of 1D-PAGE LC/MS/MS proteome: Application on angiotensin II stimulated smooth muscle cells secretome. *Mol Cell Proteomics* (2008).
17. Li, X.J., Zhang, H., Ranish, J.A. & Aebersold, R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem* 75, 6648-6657 (2003).
18. Han, D.K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* 19, 946-951 (2001).
19. MacCoss, M.J. *et al.* A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* 75, 6912-6921 (2003).
20. Silva, J.C. *et al.* Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* 5, 144-156 (2006).
21. Choi, H., Fermin, D. & Nesvizhskii, A.I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol Cell Proteomics* 7, 2373-2385 (2008).
22. Braisted, J.C. *et al.* The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* 9, 529 (2008).

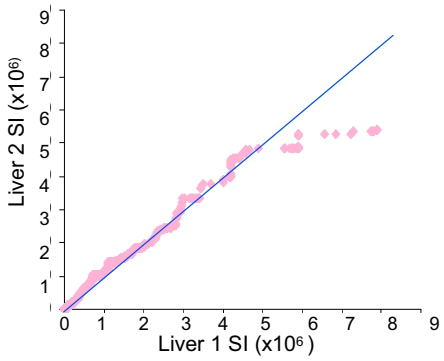
23. Lu, P. *et al.* Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25, 117-124 (2007).
24. Durr, E. *et al.* Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat Biotechnol* 22, 985-992 (2004).
25. Service, R.F. Proteomics. Proteomics ponders prime time. *Science* 321, 1758-1761 (2008).
26. Koziol, J.A., Feng, A.C. & Schnitzer, J.E. Application of capture-recapture models to estimation of protein count in MudPIT experiments. *Anal Chem* 78, 3203-3207 (2006).
27. Li, Y. *et al.* Enhancing identifications of lipid-embedded proteins by mass spectrometry for improved mapping of endothelial plasma membranes in vivo. *Mol Cell Proteomics* 8, 1219-1235 (2009).
28. Tabb, D.L., Huang, Y., Wysocki, V.H. & Yates, J.R., 3rd Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem* 76, 1243-1248 (2004).
29. Tabb, D.L. *et al.* Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* 75, 1155-1163 (2003).
30. Tabb, D.L., Fernando, C.G. & Chambers, M.C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6, 654-661 (2007).
31. Tabb, D.L. *et al.* DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* 7, 3838-3846 (2008).
32. Venable, J.D. *et al.* Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 1, 39-45 (2004).
33. Oh, P., Schnitzer, JE in Cell Biology: A Laboratory Handbook. (ed. C. J) 34-36 (Acedemic Press, Orlando; 1998).
34. Schnitzer, J.E. *et al.* Separation of caveolae from associated microdomains of GPI-anchored proteins. *Science* 269, 1435-1439 (1995).
35. Klimek, J. *et al.* The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. *J Proteome Res* 7, 96-103 (2008).
36. Forner, F. *et al.* Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* 5, 608-619 (2006).
37. Kramer, C.Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* 12, 309-310 (1956).
38. Tukey, J.W. Some selected quick and easy methods of statistical analysis. *Trans N Y Acad Sci* 16, 88-97 (1953).
39. Kendall, M. Multivariate analysis, Edn. 2nd. (MacMillan, New York; 1980).
40. Mirkin, B. Mathematical Classification and Clustering. (Kluwer Academic Publishers, 1996).
41. Cheng, Y., Church, GM. in Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology 93-1032000).
42. Hartigan, J. Direct clustering of a data matrix. *Journal of the American Statistical Association* 67, 123-129 (1972).

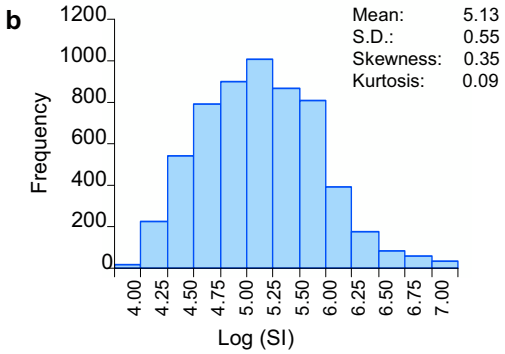
No.	abbreviation	Equation
1	SI	$SI = \sum_{k=1}^{pn}(\sum_{j=1}^{sc} i_j)_k$
2	SI_{act}	$SI_{act} = [\sum_{k=1}^{pn} (\sum_{j=1}^{sc} i_j) / \sum_{j=1}^{act} i_j$
3	SI_{MPI}	$SI_{MPI} = SI / [(\sum_{j=1}^{n} SI_j) / n]$
4	SI_{MI}	$SI_{MI} = SI / [(\sum_{j=1}^{n} SI_j) / (\sum_{j=1}^{n} SC_j)]$
5	SI_p	$SI_p = SI / (\sum_{j=1}^{PN} p_j)$
6	SI_{TSC}	$SI_{TSC} = SI / \sum_{j=1}^{n} SC_j$
7	SI_{GI}	$SI_{GI} = SI / \sum_{j=1}^{n} SI_j$
8	SI_L	$SI_L = SI / L$
9	SI_N	$SI_N = SI_{GI} / L$
10	SC_N	$SC_N = SC / [\sum_{j=1}^{n} SI_j / L]$

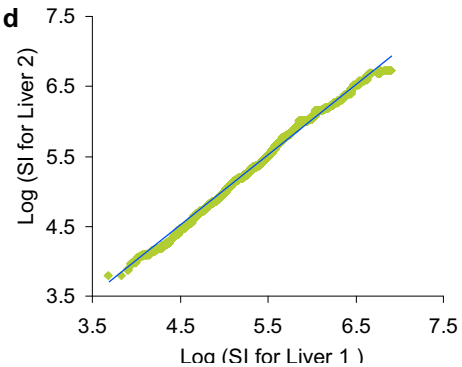
Supplementary Table 1: Summary of the normalization equations.

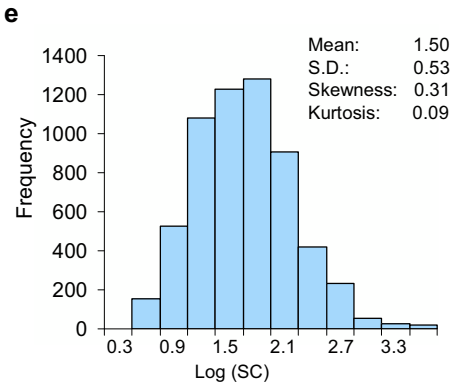
From left to right, columns contain the equation number, the abbreviation of each equation used in the paper and the equations. Equation abbreviations: spectral index (SI), spectral count (SC), peptide number (PN), peptide length (L)

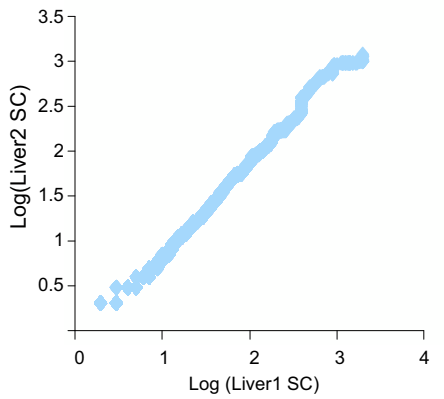

a
Mean: 332255.29
S.D.: 709818.44
Skewness: 5.74
Kurtosis: 40.28
Frequency
SI (x10⁵)

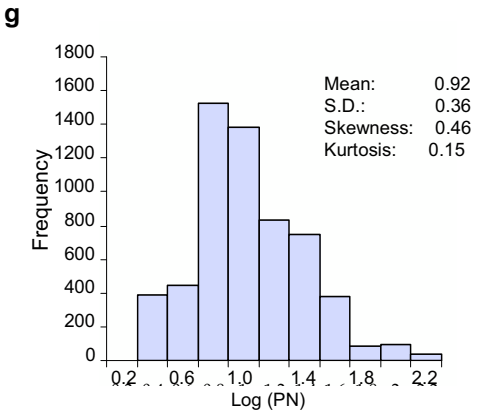

c
Liver 2 SI (x10⁶)
Liver 1 SI (x10⁶)

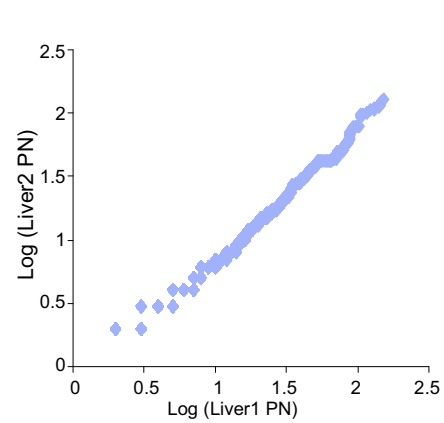

b
Mean: 5.13
S.D.: 0.55
Skewness: 0.35
Kurtosis: 0.09
Frequency
Log (SI)

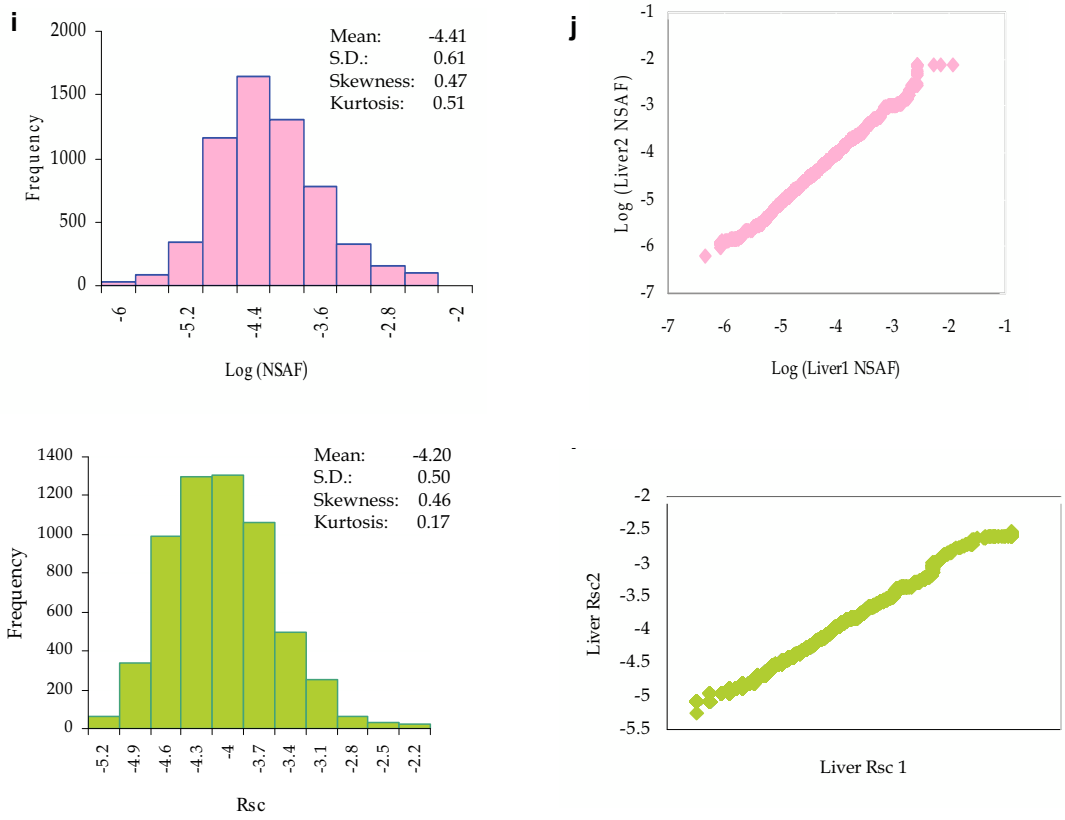

d
Log (SI for Liver 2)
Log (SI for Liver 1)


e
Mean: 1.50
S.D.: 0.53
Skewness: 0.31
Kurtosis: 0.09
Frequency
Log (SC)

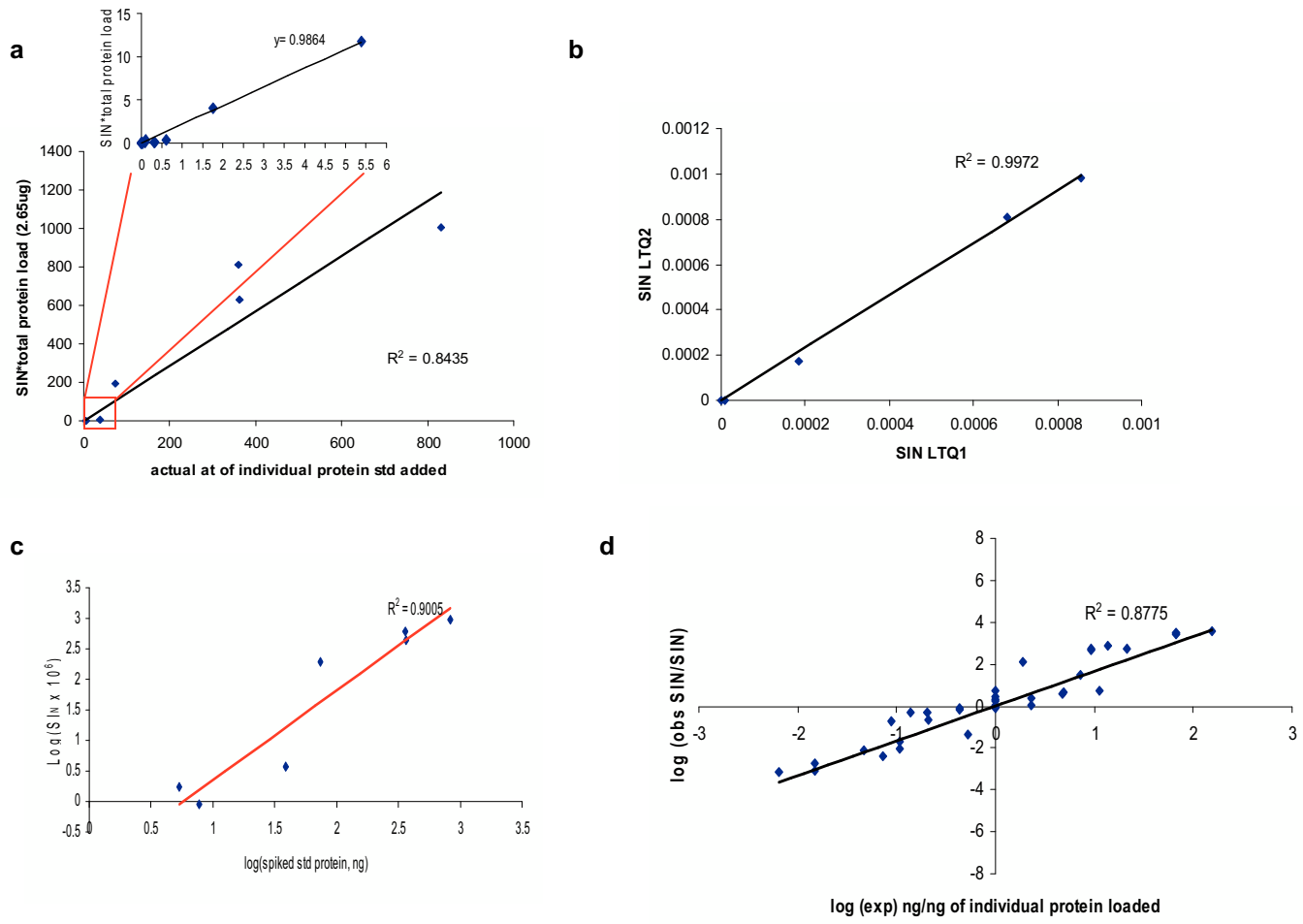

f
Log(Liver2 SC)
Log (Liver1 SC)


g
Mean: 0.92
S.D.: 0.36
Skewness: 0.46
Kurtosis: 0.15
Frequency
Log (PN)


h
Log (Liver2 PN)
Log (Liver1 PN)

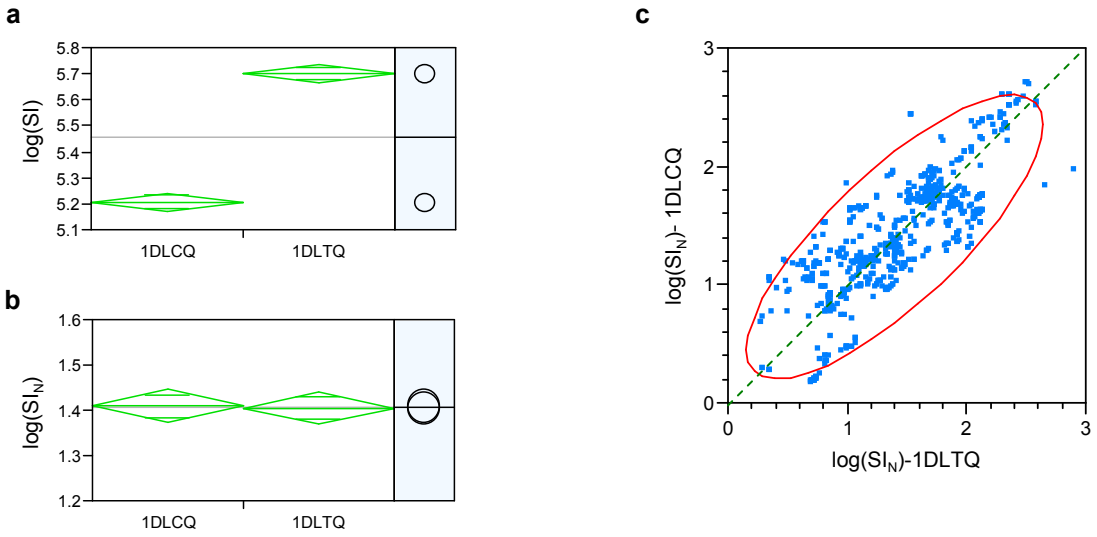


Supplementary Figure 1: Distribution of abundance features from liver datasets.
Two MS datasets from replicate liver endothelial cell plasma membrane samples were pooled and the means, standard deviations (S.D.), skewness and kurtosis are presented for SI before (**a**) and after (**b**) log transformation. To determine if the two datasets come from a common distribution, the replicate datasets were graphed using Q-Q plots with liver 1 on the y-axis and liver 2 on the x-axis. Data before (**c**) and after (**d**) log transformation is shown. A 45-degree line showing perfect correlations is plotted for reference. Similar data is plotted for **e, f**) Spectral count (SC), and **g, h**) peptide number (PN), **i, j**) NSAF normalized data, **k, l**) Rsc.



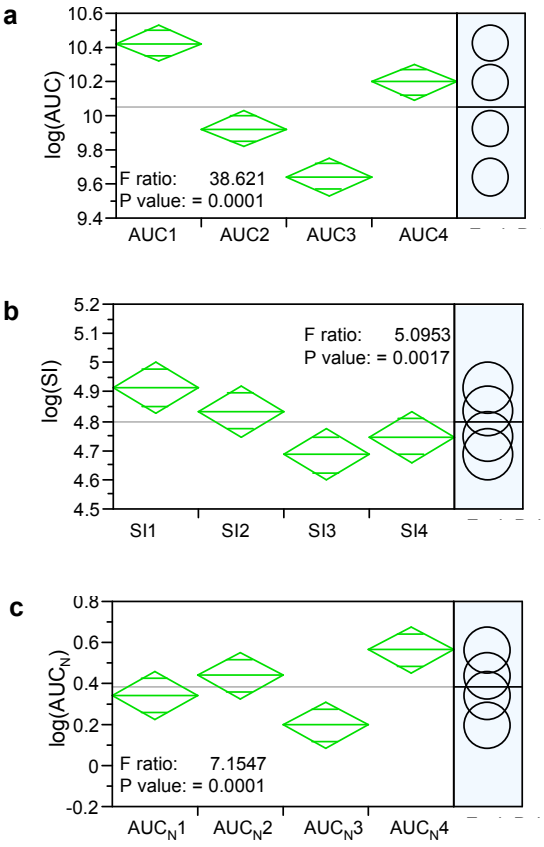
Supplementary Figure 2: Validation of SI_N as a relative quantitative tool using a protein mixture of known content

a) A protein standard mix spanning a wide dynamic range (0.5 – 50,000fmol) was analyzed by 2DLC. SI_N values for each protein were calculated and converted to observed ng amount of protein using the RPQ-P equation (see online methods) and plotted against the expected ng amount (actual ng amount of protein added to mixture). The bottom left portion of the graph was zoomed and expanded to facilitate better visualization of the actual fit for the low abundance proteins. The $R^2 = 0.8435$. b) The protein standard mixture was analyzed as (a) above but two replicates were analyzed on different LTQ mass spectrometers. The SI_N value for each protein identified on one machine was plotted against the SI_N value for the same protein identified on the other machine. R2 = 0.9972. c) SI_N values for each spiked protein were calculated, averaged values across 3 replicates and the log10 values were plotted against the log10 amount of spiked protein. d) Protein ratios were calculated for all spiked proteins by generating a ratio between the SI_N value of a specific protein divided by the SI_N value of another protein in the mixture, log10 ratios were generated for all different combinations of proteins and plotted against the log10 expected ratio (generated from actual ng value of the protein added to the mixture divided by the ng value of the other protein).



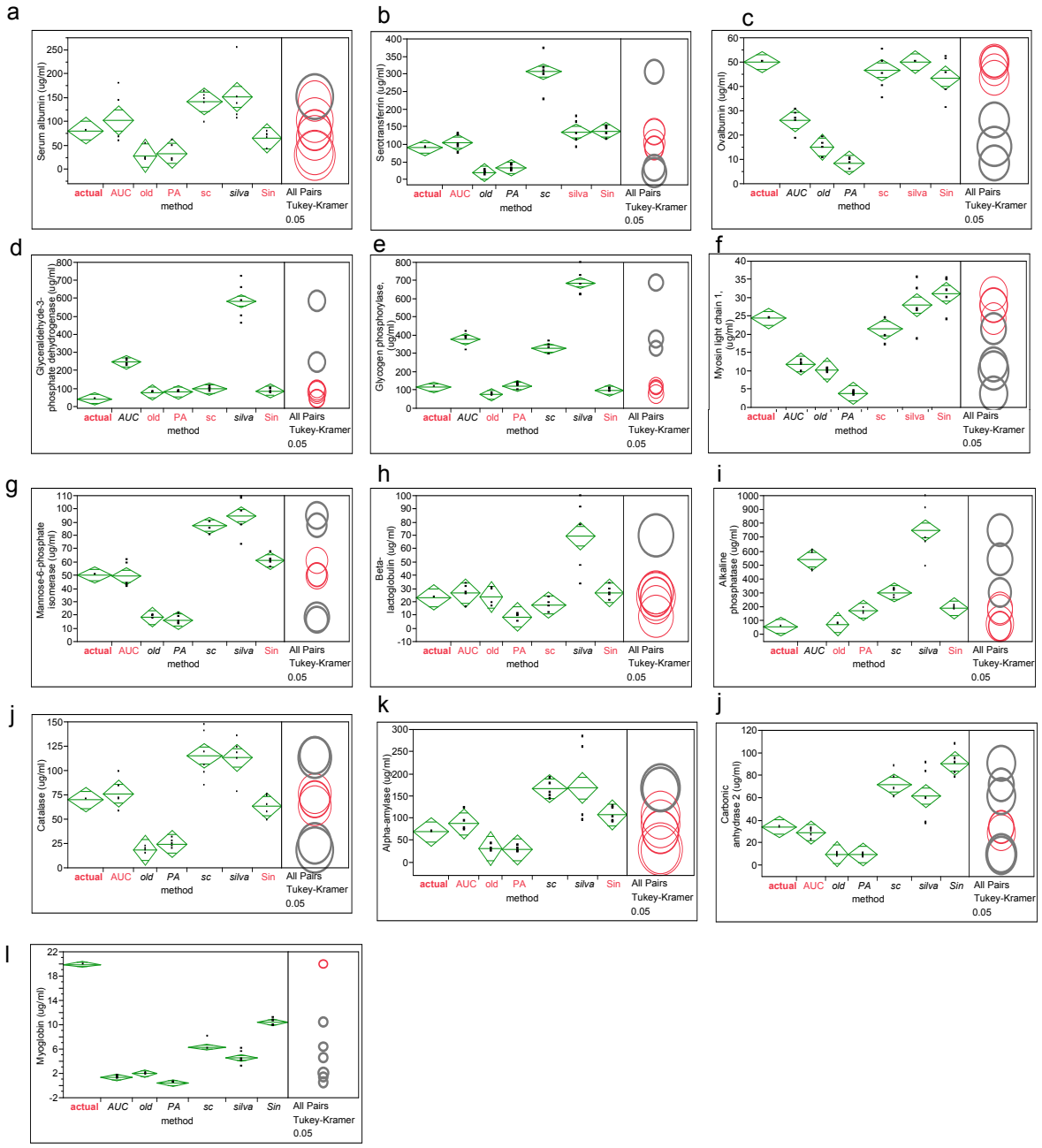
Supplementary Figure 3: Statistical analysis of the normalization methods applied to the datasets from different mass spectrometers.

Lung endothelial cell plasma membranes were separated by SDS-PAGE followed by 1D-RP-LC-MS/MS analysis of all trypsin-digested gel slices using either an LCQ (1DLCQ) or an LTQ (1DLTQ) mass spectrometer. The SI and SI_N values from the 769 proteins common between the two measurement types (3 replicates for each measurement) were averaged and plotted using the mean diamonds. The x-axis represents the 2 different MS measurement types and the y-axis represents the log of the **a**) raw SI values or **b**) SIN. **c**) A bivariate fit of the SI_N normalized 1DLCQ and 1DLTQ datasets indicates a strong positive linear correlation between the two datasets, which is confirmed by the Pearson's correlation of 0.796. The oval nature of the density ellipse also indicates the significant correlation between the datasets.



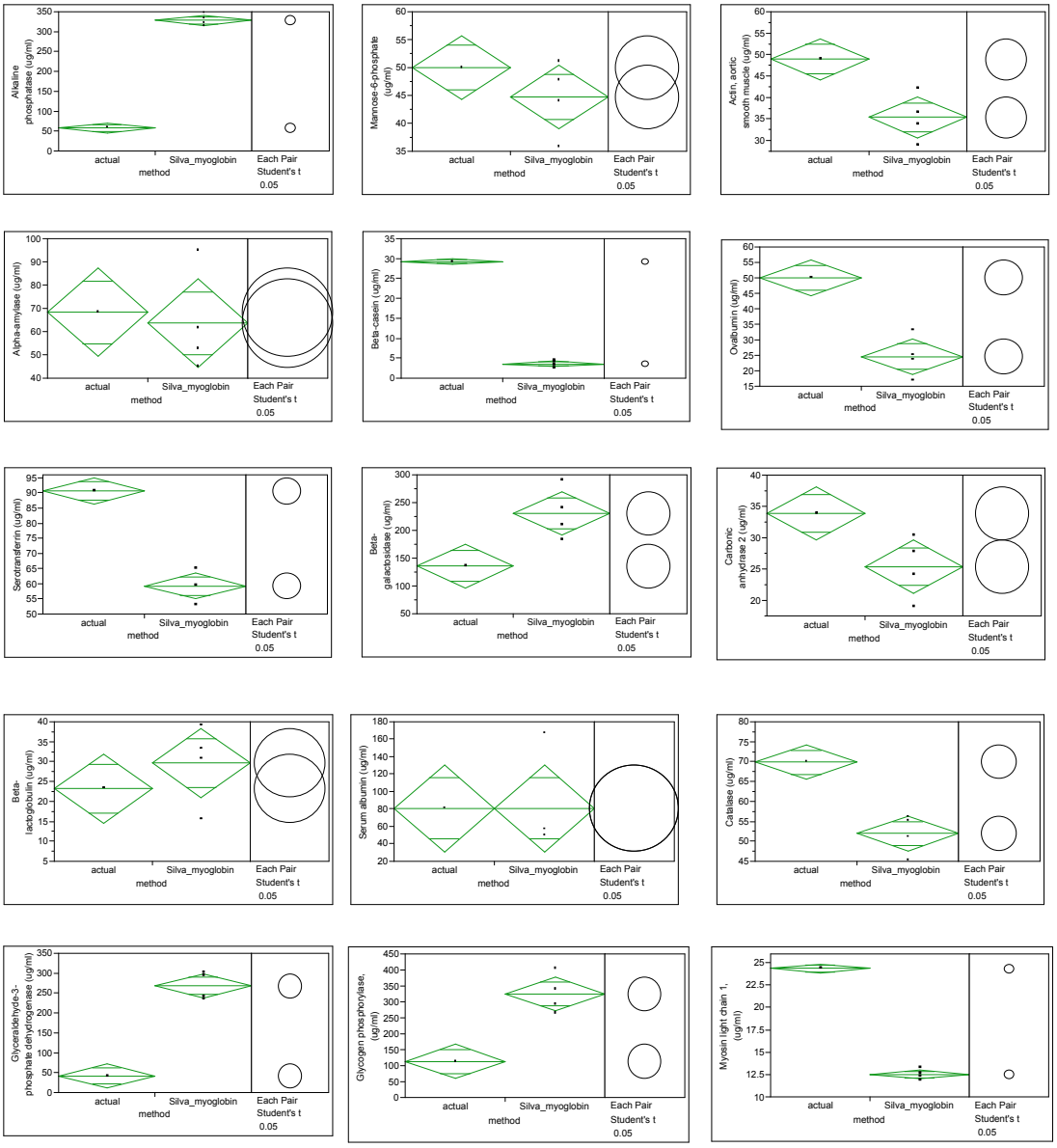
Supplementary Figure 4: Comparison of SI and Peak area (AUC) across replicate samples

Liver endothelial cell plasma membranes (4 replicates) were separated by SDS-PAGE. Gel lanes were cut into 72 slices for in-gel proteolytic digestions. Peptides extracted from each gel slice were pooled into 7 groups and then lyophilized. One out of these 7 groups (number 4) was analyzed by 2D-LC-MS/MS as described in the supplementary methods. Proteins common to all 4 replicate MS measurements were identified and AUC values for each peptide identifying these proteins were manually extracted using the AUC feature of Bioworks 3.2. SI values were also calculated. Summary statistics (mean and 95% CI) for the (a) AUC and (b) SI were plotted using the mean diamonds and comparison circles methods as described in the online methods. X-axis shows the 4 replicate measurements and the y-axis represents the log of the abundance feature being examined. For analysis of difference in mean intensities between multiple replicate samples, analysis of variance (ANOVA, one-way, $P<0.05$) was performed. (c) AUC values were normalized using the protein length and global equation as in SI_N, where the AUC replaces SI in the calculation. The normalized values were plotted and analyzed as described above.



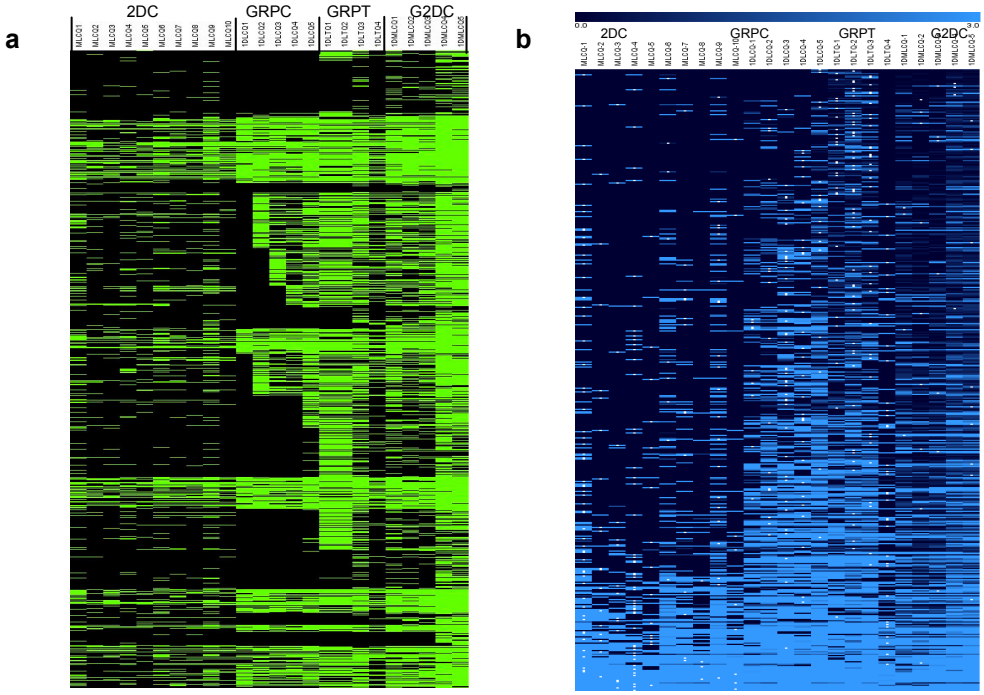
Supplementary Figure 5: Statistical analysis of the comparison of proteins quantification across replicate measurements using 6 methods (relative to known value)

The mean and 95% confidence interval (CI) for protein abundance, as determined by various relative quantitative methods, were plotted for all proteins from the “standard protein mixture” (that were detected in all 10 replicate analysis) and compared to the actual loaded amount using ANOVA, and individual means were compared using the Tukey-Kramer HSD method (methods online). Quantitative methods that were not significantly different from the actual protein abundance, as determined by overlapping mean diamonds and confidence circles and confirmed by ANOVA are highlighted in red in the Figure.



Supplementary Figure 6: **Statistical analysis comparing the Silva AUC method to the actual protein amount, where the protein with the least variation across replicates was chosen as the internal standard.**

The mean and 95% confidence interval (CI) for protein abundance, as determined by the Silva *et al* method[19], using the protein standard with the least variation across the replicates as the "spiking control", were plotted for all proteins from the "standard protein mixture" (that were detected in all 10 replicate analysis) and compared to the actual loaded amount using students t-test. If the quantitative method was not significantly different from the actual protein abundance, the mean diamonds and confidence circles overlap.



Supplementary Figure 7: **Heatmap of SIN values facilitates assessment of reproducibility between mass spectrometry methodologies.**

a) Protein detection map: SI_N was applied to 24 rat lung ECPM MS datasets from 4 different methods (see online methods Each row represents one protein and each column represents one MS experimental measurement.

A binary system was used with a green signal given to proteins positively identified versus black when not identified.

b) Protein quantification map. Protein levels were estimated by converting the SI_N value to ng values (online methods). The 500 most abundant proteins from the datasets were presented in this heat map. The data range is from 0 (dark blue) to >=3 (light blue) ng. Small white dots in the heat map indict proteins detected at levels above 3ng, thus exceeding the values represented by the color scale.